Risk Analytics in Banking: Understanding Loan Default Tendencies Based on Consumer and Loan Attributes.

By: Hossein Rajabi, Jingwen Yu, Dazhi Li, Priya Gill, Manas Agrawal

Problem

Overall Research Topic: Risk Factors Influencing Loan Default Tendencies: An In-depth Analysis of Consumer Attributes, Loan Characteristics, and Previous Loan History

Subtopic 1 - Consumer Attributes:

- Do certain consumer demographics (e.g., age, education level, employment type) have a higher likelihood of defaulting on their loans than others?
- How does the financial stability of a consumer (e.g., income level, existing debts) impact their loan repayment behavior?

Subtopic 2 - Loan Characteristics

- What are the most common types of loans taken by consumers?
  (Ex: Personal loans, Mortgages, Education loans, etc.)
- Are there specific loan characteristics (e.g., loan duration, interest rate, loan amount) that correlate with higher default rates?
  (Ex: Larger loan amounts, Higher interest rates, Longer loan durations, etc.)

Subtopic 3 - Previous Loan History

- Does a consumer's past loan repayment behavior predict their future loan default tendencies?
- Can we spot patterns in consumers who have defaulted in the past versus those who haven't, based on their interaction with previous loans?
  (Ex: Multiple short-term loans, Early loan repayments, Previous loan defaults, etc.)

Impact

Consumer Attributes and Loan Defaults

The analysis of consumer attributes and loan defaults directly highlights how specific consumer attributes are related to the likelihood of loan defaults. It also equips lending institutions with valuable tools and knowledge about which attributes should be monitored closely, ultimately leading to more effective loan risk management. Additionally, it provides insights to potential borrowers, by offering guidance on how to improve and enhance their chances of loan approval based on certain attributes.

Loan Attributes Influencing Default Tendency

By analyzing the loan attributes that influence default tendency, we can provide a clear and comprehensive understanding of the specific loan conditions or features that can contribute to higher default rates. Additionally, it empowers financial institutions to tailor their loan offerings in a way that minimizes default risks while optimizing profitability at the same time. Furthermore, it helps consumers by alerting them to common loan terms that are associated with defaults, thereby guiding them in making well-informed decisions when borrowing.

Effect of Past Loan Actions on Today's Choices

Knowing how a person acted with past loans helps banks decide on new loan applications. This information makes it easier to judge if a loan might not be paid back. Recognizing trends from past actions is good for both banks and borrowers. Banks can make smarter choices, and borrowers can understand how their past choices might affect new loans.

Data

Application File

For our analysis, the primary dataset, 'application_data.csv', provides insights into the client's information at the time they apply for a loan. This dataset, referred to as the Main Application Record DataFrame, is substantial in size with 15,523 rows and 122 columns. It's diverse in content with 95 columns recorded as float64, 11 columns as int64, and 16 columns as object types. However, some data points have raised concerns. Features like AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, and AMT_REQ_CREDIT_BUREAU_YEAR are missing 2,085 entries each. Furthermore, certain descriptive statistics look strange. For instance, the mean value for DAYS_BIRTH being -15,996.18 days, and the minimum DAYS_EMPLOYED being -15632.00 days, seems odd and needs more checking.

This data is expansive, offering a window into essential details like the applicant's income and family status. This allows us to delve into questions like the client's capability to repay loans based on their financial standings. We need to be cautious though, as potential biases, like inconsistent data recording methods, could change our results. Within the 'application_data.csv', we've noticed some gaps, such as missing occupation types for certain applicants. The ambiguity raises questions: Are these people unemployed or was the data simply not recorded? To navigate this, we intend to shift our

focus to analyzing clients based on their yearly earnings. This will help in figuring out patterns, for instance, if individuals with higher annual salaries show a lower tendency for loan defaults, or if those with limited income are more inclined towards it.

Previous Application File

Moving on, the 'previous_application.csv' file provides the client's past loan history. This dataset, known as the Previous Application Record DataFrame, consists of 34,180 rows and 37 columns. Within these, 18 columns are float64, 3 columns are int64, and 16 columns are object types. There are noticeable missing data in columns like RATE_DOWN_PAYMENT and RATE_INTEREST_PRIMARY with 17,361 and 34,075 missing values, respectively. Also, we've spotted statistics that appear off. The mean for DAYS_DECISION being -900.634132 and the minimum for DAYS_FIRST_DUE being -2891.000000 days require deeper investigation.

Columns File

Another dataset we have is the 'columns_description.csv'. Known as the Descriptions DataFrame, it's more concise, having 160 rows and 5 columns. It's structured with 1 column of int64 and 4 columns of object types. An area of attention here is the 'Special' column, which has 91 missing entries. For clarity, the 'columns_description.csv' will serve as our guide, explaining the variables in our datasets. This will be helpful in understanding the data clearly and making sure we read it right. In our first look, it's good to see that there aren't any duplicate values in any of the three data sets.

Nice-to-Have Variables

Credit Score: Helps us gauge the likelihood of a client repaying. A higher score often means lower default risk.

Loan Purpose: Allows us to assess the potential risk tied to the reason for borrowing. Some reasons might be associated with higher default rates.

Co-applicant/Guarantor Info: Can give us more confidence in loan approvals, knowing there's backup for repayment.

Methods -- Variables and Visualizations

New Variables Needed

To improve our analysis, we'll introduce the following variables:

1. AGE: Derived from DAYS_BIRTH.
2. AGE_CATEGORY: Using AGE, this categorizes individuals into age groups.
3. EMPLOYMENT_DURATION: Based on DAYS_EMPLOYED.
4. LOAN_DURATION: Extracted from DAYS_LAST_DUE and DAYS_FIRST_DUE.

Subtopic 1 - Consumer Attributes

Relevant Variables:

The variables that are relevant to this subtopic are CODE_GENDER, AGE, AGE_CATEGORY, NAME_EDUCATION_TYPE, NAME_INCOME_TYPE, OCCUPATION_TYPE, CNT_CHILDREN, AMT_INCOME_TOTAL, and EMPLOYMENT_DURATION.

Visualizations:

- Bar charts: Male vs. female default rates.
- Histogram: Age distribution of clients.
- Bar charts: Default rate across education types and income types.
- Scatter plots: Number of children vs default rate, and employment duration vs default tendency.

Analyzing Relationships:

To analyze relationships, we can use a box plot to compare how income varies with different education types. We can also use correlation matrices for understanding the relationship between income and child count.

Subtopic 2 - Loan Characteristics

Relevant Variables:

The variables that are relevant to this subtopic are NAME_CONTRACT_TYPE, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, REGION_POPULATION_RELATIVE, and LOAN_DURATION.

Visualizations:

- Histograms: Distribution of contract types and credit amounts.

- Scatter plots: Relationship between annuity amount vs default rate, goods price vs default rate, and LOAN_DURATION against default rate.
- Heatmap: Default rate across regions based on population.

Analyzing Relationships:
To analyze relationships, we can create correlation matrices to observe the correlation between credit amount and contract type. We can also create a pair plot to determine the interrelationship of loan variables and contract types.

Subtopic 3 - Previous Loan History

Relevant Variables:
The variables that are relevant to this subtopic are NAME_CONTRACT_STATUS (Approved, Canceled, Refused, Unused offer), AMT_APPLICATION, AMT_CREDIT, and DAYS_DECISION.

Visualizations:
- Bar chart: Distribution of previous loan statuses.
- Scatter plots: Previous application amount vs default rate and previous credit amount vs default rate.
- Histogram: Days since the previous decision.

Analyzing Relationships:
To analyze relationships in regards to this subtopic, we can create a Box plot to determine the variation of AMT_APPLICATION and AMT_CREDIT, based on the NAME_CONTRACT_STATUS.

Deep Dive into One Visualization
Focusing on the scatter plot comparing AMT_ANNUITY with the default rate:
- *Utility:* Assesses if higher monthly payments correlate with higher default rates.
- *X-axis:* Monthly loan payment (AMT_ANNUITY).
- *Y-axis:* Default rate.
- *Interpretation:*
  An upward trend suggests an increase in default rates with rising monthly payments.
- *Business Implication:*
  Adjusting interest rates or offering flexible payment options may be considered based on findings.

Concerns

Concerns about the Integrity of the Dataset:

Missing Data: The presence of missing entries in features like AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, and others, reduces the completeness of the dataset. This might lead to skewed analysis since there's information absence.

Odd Data Entries: Some descriptive statistics, like the mean value for DAYS_BIRTH and DAYS_EMPLOYED, appear unusual. Such inconsistencies could be data entry errors, which can distort the results.

Ambiguities: There are gaps in the 'application_data.csv' dataset, for instance, the missing occupation types for certain applicants. This kind of missing information can introduce ambiguity to the analysis.

Previous Loan Data: The 'previous_application.csv' has columns with substantial missing data, which can affect the depth and accuracy of analysis related to previous loan history.

Potential Drawbacks of Analysis and Visualizations Proposed:

Consumer Demographics: While analyzing demographics like age, education level, etc., there's a possibility of introducing biases or making over-generalizations about specific groups, which may not capture the nuanced nature of loan defaults.

Interpreting Loan Characteristics: The analysis attempts to correlate loan characteristics like loan amount, duration, and interest rate with default rates. However, correlation does not imply causation. Just because two variables move together doesn't mean one caused the other.

Previous Loan History as Predictor: Relying heavily on previous loan history might overlook other potential factors influencing loan repayment behavior.

Visualization Limitations: While visualizations like scatter plots and histograms provide insights, they might not capture the multifaceted relationships among variables. There's also a risk of misinterpreting the visual data or relying too heavily on them without considering other factors.

Additional Concerns about the Overall Analysis:

Narrow Focus: The analysis mainly focuses on certain variables and may overlook others, which could be just as significant. This tunnel vision approach might miss other potential red flags or indicators.

Data Quality: Without a clear understanding of how the data was collected, there's potential risk concerning the dataset's accuracy and reliability.

Subjectivity: The categorization of newly introduced variables like AGE_CATEGORY might introduce subjective biases. How these categories are defined could sway the results in one direction or another.

External Factors: The analysis does not take into account external economic factors, which can greatly influence loan default rates, such as economic downturns, job market fluctuations, or other macroeconomic indicators.

Future Predictability: Just because certain patterns exist in the past does not mean they'll continue in the same way in the future. Relying on past data to predict future behavior has its limitations.